Data Wrangling (daw)
HS 2020
Cédric Huwyler

Fachhochschule Nordwestschweiz

Prüfung

Nachname	Vorname

Hinweise

Datum, Zeit, Ort: 11. Februar 2021, 8:15 Uhr, Zimmer 1.023

Prüfungsdauer: 90 Minuten

Rahmenbedingungen

- Sie bestreiten die Prüfung auf Ihrem eigenen Computer und geben am Schluss diese Prüfungsblätter und ein begleitendes Jupyter-Notebook (an `cedric.huwyler@fhnw.ch`) ab.
- Benötigte Datensets werden vom Dozierenden zum Start der Prüfung zur Verfügung gestellt.
- Für die maximale Punktzahl wird ein sauber dokumentierter Lösungsweg im Notebook und/oder auf dem Prüfungsblatt erwartet.
- Die Prüfung ist Open-Book und Sie dürfen das Internet zur Informationssuche benutzen.
- Kommunikation mit anderen Studierenden oder aussenstehenden Personen während der Prüfung ist nicht erlaubt. Zuwiderhandlung zieht die Note 1 mit sich.

Benotung

Aufgabe	1	2	3	Total	Note
Maximale Punktzahl	27	27	21	75	
Erreichte Punktzahl					

Aufgabe 1. (27 Punkte)

In der Datei `'tips.csv'` finden Sie einen Datensatz mit Erhebungen einer Bar. Darin enthalten sind pro Tisch: totaler Rechnungsbetrag, Trinkgeldbetrag, Angabe Raucher/Nicht-Raucher, Wochentag, Tageszeit und Gruppengrösse.

(a) **(1 Punkt)** Importieren Sie das Datenset in ein Data Frame.

(b) **(6 Punkte)** Beschreiben Sie die Datentypen der verschiedenen Merkmale: Welche sind diskret und welche stetig? Welche der Merkmale davon sind nominal-, ordinal-, intervall- und welche verhältnisskaliert?

(c) **(2 Punkte)** Geben Sie ein Beispiel eines stetigen, aber nicht verhältnisskalierten Merkmals an und erklären Sie kurz:

(d) **(6 Punkte)** Zurück zum erstellten Data Frame. Formatieren Sie die diskreten Variablen wo nötig als *kategorische* Variablen. Dabei soll die Variable im resultierenden Data Frame über die verschiedenen Ausprägungen des Merkmals Bescheid wissen und wo möglich auch die Bedeutung der Ordnungsrelationen $<$ und $>$ verstehen können.

(e) **(6 Punkte)** An welchem Wochentag nahm die Bar durchschnittlich prozentual am meisten Trinkgeld ein? Geben Sie ein informatives Data Frame aus und unterlegen Sie Ihre Antwort mit einem Barplot. Die Wochentage sollen dabei in der korrekten Reihenfolge angeordnet sein.

(f) **(6 Punkte)** Wie hängt die prozentuale Trinkgeldhöhe qualitativ mit der Gruppengrösse zusammen? Beurteilen Sie mit einem Boxplot der Trinkgeldhöhe pro Gruppengrösse (alle Boxplots in einem Plot). Geben Sie zur Sicherheit auch die Verteilung der Gruppengrössen an - spielt diese bei Ihrer Beurteilung eine Rolle?

Aufgabe 2. (27 Punkte)

In den zur Verfügung gestellten Dateien finden Sie die Excel-Datei `'diabetes.xlsx'` mit einem Datenset und einer Beschreibung dazu in den verschiedenen Sheets.

(a) **(1 Punkte)** Lesen Sie die Beschreibung und importieren Sie das Datenset in ein Data Frame.

(b) **(7 Punkte)** Enthält das Datenset fehlende Werte? Untersuchen Sie die Wertebereiche der einzelnen Spalten genau. Stellen Sie sicher, dass alle klar erkennbaren fehlenden Werte mit `NaN` auch als solche gekennzeichnet sind. Geben Sie die Anzahl der fehlenden Werte pro Spalte absolut und in Prozent aus.

(c) **(1 Punkt)** Erklären Sie kurz, warum wir hier das komplette Entfernen der Samples mit fehlenden Werten vermeiden möchten und lieber eine passende Imputationsstrategie wählen.

(d) **(7 Punkte)** Erstellen Sie ein neues Data Frame und wenden Sie folgenden Strategien an:

- `gtt` und `triceps_skin_fold_thickness`: Imputation mit Median
- `blood_pressure` und `bmi`: Imputation mit Durchschnitt

Stellen Sie am Schluss sicher, dass im neuen Data Frame keine fehlenden Werte mehr vorkommen.

(e) **(4 Punkte)**

Alternativ können fehlende Werte auch modellbasiert ersetzt werden. Dazu soll ein *k nearest neighbours* (KNN) - Modell verwendet werden (mit `n_neighbours`=5). Verwenden Sie dazu die Klasse `sklearn.impute.KNNImputer` von Scikit-Learn.

Stellen Sie auch hier sicher, dass im neuen Data Frame keine fehlenden Werte mehr vorkommen.

Hinweis: Natürlich sollte die Variable `class` nicht als Grundlage zur Imputation benutzt werden, da sonst Information darüber in die unabhängigen Variablen leakt.

(f) **(7 Punkte)** Wir möchten zum Schluss auf dem imputierten Datenset aus Teilaufgabe (e) untersuchen, ob und wie stark Fettleibigkeit zum Diabetesrisiko beiträgt. Erstellen Sie dazu im Data Frame eine neue Spalte `'bmi_class'` mit der folgenden Einteilung:

$$\texttt{bmi\_class} = \begin{cases} \texttt{'underweight'}, & \texttt{bmi} < 18.5 \\ \texttt{'normal'}, & 18.5 \leq \texttt{bmi} < 25 \\ \texttt{'overweight'}, & 25 \leq \texttt{bmi} < 30 \\ \texttt{'obese'}, & \texttt{bmi} \geq 30 \end{cases}$$

Die Spalte soll über die Ordinalskala der BMI-Klasse informiert sein.

Berechnen Sie nun das mittlere Diabetesrisiko und die Anzahl der Probanden pro BMI-Klasse (alles in einem Data Frame). Was ist Ihre Schlussfolgerung?

Aufgabe 3. (21 Punkte)

Das *Gapminder*-Datenset enthält Daten zum Bruttoinlandprodukt (BIP) pro Kopf (engl. GDP per capita), zur Lebenserwartung (eng. life expectancy) und zur Bevölkerungsgrösse (engl. population) von 142 Ländern auf 5 Kontinenten.

In `'gapminder.json'` finden Sie ein verschachteltes JSON-Datenset. Dabei ist das oberste Hierarchielevel der Kontinent, das zweitoberste das Land und die relevanten Grössen finden sich schliesslich in der dritten Hierarchiestufe.

(a) **(9 Punkte)** Lesen Sie das Json-File als Dictionary ein und bringen Sie es in die Form eines flachen Data Frames:

	continent	country	gdpPercap_1952	gdpPercap_1957	gdpPercap_1962	...	pop_1987	pop_1992	pop_1997	pop_2002	pop_2007
0	Africa	Algeria	2449.008185	3013.976023	2550.816880	...	23254956.0	26298373.0	29072015.0	31287142	33333216
1	Africa	Angola	3520.610273	3827.940465	4269.276742	...	7874230.0	8735988.0	9875024.0	10866106	12420476
2	Africa	Benin	1062.752200	959.601080	949.499064	...	4243788.0	4981671.0	6066080.0	7026113	8078314
3	Africa	Botswana	851.241141	918.232535	983.653976	...	1151184.0	1342614.0	1536536.0	1630347	1639131
4	Africa	Burkina Faso	543.255241	617.183465	722.512021	...	7586551.0	8878303.0	10352843.0	12251209	14326203

Hinweis: Das geht zum Beispiel mit zwei verschachtelten `for`-Loops.

(b) **(9 Punkte)** Bringen Sie nun das Data Frame in die folgende Form:

	continent	country	year	gdpPercap	lifeExp	pop
0	Africa	Algeria	1952	2449.008185	43.077	9279525.0
1	Africa	Algeria	1957	3013.976023	45.685	10270856.0
2	Africa	Algeria	1962	2550.816880	48.303	11000948.0
3	Africa	Algeria	1967	3246.991771	51.407	12760499.0
4	Africa	Algeria	1972	4182.663766	54.518	14760787.0

(c) **(3 Punkte)** Stellen Sie die Entwicklung der durchschnittlichen Lebenserwartung pro Kontinent graphisch dar. Gewichten Sie in der Berechnung des Durchschnitts nach Bevölkerungsgrösse pro Land.